# Usability evaluation of a virtual museum interface

Article (Unspecified)

INFORMATICA, 2006, Vol. 17, No. 3, 1–17



# Usability Evaluation of a Virtual Museum Interface

Athanasis KAROULIS
*Aristotle University of Thessaloniki*
*PO Box 888 - 54124 Thessaloniki, Greece*
*e-mail: karoulis@csd.auth.gr*

Stella SYLAIOU
*Aristotle University of Thessaloniki*
*PO Box 888 - 54124 Thessaloniki, Greece*
*University of Sussex, U.K.*
*e-mail: sylaiou@photo.topo.auth.gr*

Martin WHITE
*University of Sussex*
*Falmer - Brighton BN1 9QH, U.K.*
*e-mail: M.White@sussex.ac.uk*

Received: April 2006

**Abstract.** The Augmented Representation of Cultural Objects (ARCO) system provides software and interface tools to museum curators to develop virtual museum exhibitions, as well as a virtual environment for museum visitors over the World Wide Web or in informative kiosks. The main purpose of the system is to offer an enhanced educative and entertaining experience to virtual museum visitors. In order to assess the usability of the system, two approaches have been employed: a questionnaire based survey and a Cognitive Walkthrough session. Both approaches employed expert evaluators, such as domain experts and usability experts. The result of this study shows a fair performance of the followed approach, as regards the consumed time, financial and other resources, as a great deal of usability problems has been uncovered and many aspects of the system have been investigated. The knowledge gathered aims at creating a conceptual framework for diagnose usability problems in systems in the area of Virtual Cultural Heritage.

**Key words:** usability evaluation, museum interface, augmented reality, cognitive walkthrough.

## Introduction

Museum collections are the source from which the museum's unique role in the cultural fabric of society emanates via their contribution to scholarship, being the instruments of its education role, and the cause of its public enlightenment (Perrot, 1977). However, large collections and certain of the artefacts they hold, remain in storage places due to the museums' lack of space, the high cost of maintaining the exhibits and the fragility of certain cultural artefacts. Current research (Jones and Christal, 2002; Scali *et al.*, 2002) and

an extensive survey to European museum sector have shown (Tsapatori, 2003) that technologies such as the World Wide Web enhanced by 3D visualization tools can provide solutions to the aforementioned problems. In addition to these, the use and integration of the promising Virtual Reality (VR), Augmented Reality (AR) and Web3D technologies in conjunction with database technology may facilitate the preservation, dissemination and presentation of cultural artefacts in museums' collections and also educate in an innovative and attractive way the wide public. *Virtual Reality* signifies a synthetic world, whereas *Augmented Reality* signifies computer generated 2D or 3D virtual worlds superimposed on the real world. *Web3D* is used to represent the application of XML (eXtended Markup Language) and VRML (Virtual Reality Markup Language) technologies to deliver interactive 3D virtual objects in 3D virtual museums (Liarokapis *et al.*, 2004). Previous research has made use of 3D multimedia tools in order to record, reconstruct and visualize archaeological ruins using computer graphics (Cosmas *et al.*, 2001) and also provides interactive AR guides for the visualization of cultural heritage sites information (Gleue and Dähne, 2001). Moreover, relevant research has demonstrated that 3D technology '*offers museums rich opportunities in a range of areas from public access to conservation*' (Shaw *et al.*, 2004). These new emerging technologies are used not only because of their popularity, but also because they provide an enhanced experience to the virtual visitors. Additionally, these technologies offer an innovative, appealing and cost-effective way of presenting cultural information. Virtual museum exhibitions can present the digitized information of cultural objects, either in a museum environment (e.g., in interactive kiosks), or through the World Wide Web.

In order to address these problems, the ARCO (Augmented Representation of Cultural Objects) (ARCO, 2004) system has been developed and described in detail in (Wojciechowski *et al.*, 2004). In this paper we report on the usability evaluation of the two main components of the ARCO system, namely the ACMA (ARCO Content Management Application) and the ARIF (Augmented Reality InterFace) subsystems.

## The ARCO System

The ARCO system allows museum curators to build, manage, archive and present virtual museum exhibitions based on 3D models of artifacts. ARCO also allows end-users to explore virtual exhibitions implemented using the system (Wojciechowski *et al.*, 2004) (Figs. 1, 2).

The cultural artifacts are digitized by means of a custom built stereo photogrammetry system (Object Modeler), mainly for digitizing small and medium sized objects and a custom modeling framework (Interactive Model Refinement and Rendering tool) that is used, in order to refine the digitized artifact. These technologies are described in detail in (Patel *et al.*, 2003). The 3D models are accompanied by images, texts, metadata information, sounds and movies. These virtual reconstructions (3D models and accompanying data sets) are represented as eXtensible Markup Language (XML) based data to allow interoperable exchange between ARCO and external heritage systems (Wojciechowski *et*

Fig. 1. Museum exhibition using VR.

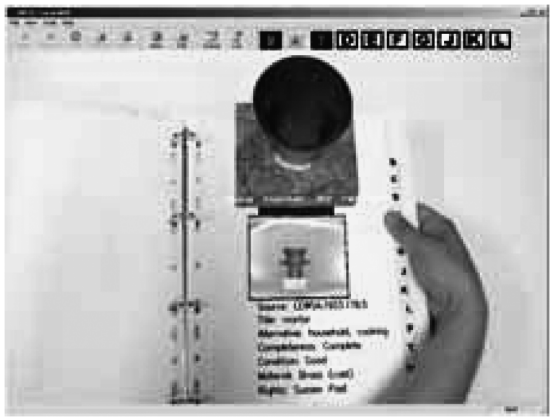

Fig. 2. Museum exhibition using AR.

*al.*, 2004). These virtual reconstructions are stored in an Oracle9i database system and managed through the use of a specially designed ARCO Content Management Application, which also allows the museum to build and publish virtual exhibitions to the Internet or a museum kiosk system.

The ARCO system is a complete tool that enables archiving of both content and context of museum objects. The interactive techniques offered can transform the museum visitors '*from passive viewers and readers into active actors and players*' (ibid).

## The ARCO Components

Two main components of the ARCO system were of interest for the evaluation: the ARCO Content Management Application (ACMA) and Augmented Reality Interface (ARIF). ACMA allows publishing of virtual museums to both Web (Fig. 1) and a specially designed application (ARIF) for switching between the Web and an AR system (Fig. 2).

The ACMA application is implemented in Java and it includes the database of the representations of cultural objects and their associated media objects, such as images, 3D models, texts, movies, sounds and relevant metadata (Mourkoussis *et al.*, 2003). It enables user-friendly management of different types of data stored in the ARCO database, through various managers, such as the *Cultural Object Manager* (deals with virtual representations of cultural artefacts), the *Presentation Manager* (manages virtual exhibitions with the help of templates) and the *Template Manager* (stores these visualization templates).

The ARIF component is a presentation or visualisation framework that consists of three main subcomponents:

- The *ARIF Exhibition Server*. Data stored in the ARCO Database is visualized on user interfaces via the ARIF Exhibition Server.
- The *ARIF Presentation Domains* with implemented web browser functionality, suited for web-based presentations.
- The *ARIF AR – Augmented reality functionality*. This sub-component provides an AR based virtual museum exhibition experience on a touch screen in the museum environment using table-top AR learning experiences, e.g., AR quizzes and on-line museum exhibitions.

## Usability Evaluation

### *Definitions*

What is *usability*? According to ISO-9241 (Ergonomic requirements for office work with visual display terminals) (ISO, 1998) standard, *usability of a system is its ability to function effectively and efficiently, while providing subjective satisfaction to its users.*

Usability of an interface is usually associated with five parameters (ISO, 1998; Nielsen, 1993), derived directly from this definition:

- *easy to learn*: the user can get work done quickly with the system,
- *efficient to use*: once the user has learnt the system, a high level of productivity is possible,
- *easy to remember*: the casual user is able to return to using the system after some period without having to learn everything all over again,
- *few errors*: users do not make many errors during the use of the system or if they do so they can easily recover them,
- *pleasant to use*: users are subjectively satisfied by using the system; they like it.

Two important conceptions regarding the usability of an interface are “transparency” and “intuitiveness” (Nielsen, 1993; Preece *et al.*, 1994). Transparency refers to the ability of the interface to fade out in the background, allowing the user to concentrate during his work on *what* he wants to do and not on *how* to do it, in our case not interfering with the learning procedure, while intuitiveness refers to its ability to guide the user through it by the use of proper metaphors and successful mapping to the real world, e.g., by

providing him with the appropriate icons, correct labeling, exact phrasing, constructive feedback etc.

Interface evaluation of a software system is a procedure intended to identify and propose solutions for usability problems caused by the specific software design. The term "evaluation" generally refers to the process of "gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context" (Preece *et al.*, 1994, p.602). A usability problem may be defined as "anything that interferes with user's ability to efficiently and effectively complete tasks" (Karat *et al.*, 1992).

There are two major evaluation approaches: *formative* and *summative* evaluation (Scriven, 1976). The former is conducted during the design and construction phase, while the latter is conducted after the product has reached the end user. The results and conclusions of the former are used mainly for bug-fixing and improving the characteristics of the interface (detecting problems and shortcomings), while the results and conclusions of the latter are used to improve the interface as a whole and meet more user needs in a following upgrade.

*Expert-based vs User-based (Empirical) Evaluation*

The most applied methodologies are the expert-based and the empirical (user-based) evaluation. Expert evaluation is a relatively cheap and efficient formative evaluation method applied even on system prototypes or design specifications up to the almost ready-to-ship product. The main idea is to present the tasks supported by the interface to an interdisciplinary group of experts who will take the part of would be users and try to identify possible deficiencies in the interface design.

However, according to (Lewis and Rieman, 1994) "you can't really tell how good or bad your interface is going to be without getting people to use it". This phrase expresses the broad belief that user-testing is inevitable in order to assess an interface. Why then, don't we use absolutely empirical evaluations but research other approaches as well? As we may see further on, the efficiency of these methods is strongly diminished by the required resources and by some other disadvantages they provide, while, on the other hand expert-based approaches have meanwhile matured enough to provide a good alternative.

The first main disadvantage of the empirical studies is the personal bias of the subjects. It is important to understand that test users can't tell you everything you might like to know, and that some of what they will tell you is useless. This is not done on purpose; for different reasons users often can not give any reasonable explanation for what happened, or why they acted in a certain way. Psychologists have done some interesting studies on these points.

Maier (1931) had people try to solve the problem of tying together two strings that hung down from the ceiling too far apart to be grabbed at the same time. One solution was to tie some kind of weight to one of the strings, set it swinging, grab the other string, and then wait for the swinging string to come close enough to reach. It's a hard problem, and few people came up with this or any other solution. Sometimes, when people were

working, Maier would "accidentally" brush against one of the strings and set it in motion. The data showed that when he did this, people were much more likely to find the solution. The point of interest for us is, what did these people say when Maier asked them how they solved the problem? They did NOT say, "When you brushed against the string that gave me the idea of making the string swing and solving the problem that way," even though Maier knew that's what really happened. So they could not and did not tell him what feature of the situation really helped them solve the problem.

Lewis and Rieman (1994) give the three prerequisites for an empirical evaluation: People, if possible real users in real circumstances; some tasks for them to perform, and; some version of the system to work with. At this point we already have another obstacle regarding the empirical approaches: All these prerequisites are required simultaneously.

On the other hand, according to Reeves (1993), expert-based evaluations are perhaps the most applied evaluation strategy. This happens mainly because they provide a crucial advantage which makes them more affordable compared to the empirical ones: it is in general easier and cheaper to find out experts eager to perform the evaluation than users. The main idea is that experts from different cognitive domains, anyway at least one from the domain of HCI and at least one from the cognitive domain under evaluation, are asked to judge the interface, everyone from his own point of view. It is important that they are all experienced, so they can see the interface through the eyes of the user and reveal problems and deficiencies of the interface. One strong advantage of the methods is that they can be applied very early in the design cycle, even on paper mock-ups. The expert's expertise allows him to understand the functionality of the system under construction, even if he lacks the whole picture of the product. A first look at the basic characteristics would be sufficient for an expert. On the other hand, user-based evaluations can be applied only after the product has reached a certain level of completion.

## Evaluation of the ARCO System

The ARCO System has been evaluated by utilizing a variety of methods, both empirical and expert-based, and some preliminary results have already been reported in (Sylaiou *et al.*, 2004). However, this study focuses only on the usability evaluation of the system and bases on two evaluation sessions, one questionnaire based and one session of Cognitive Walkthrough. The questionnaire based session was performed by the museum curators and assessed the ACMA as well the ARIF interface. The Cognitive Walkthrough was performed by "visitors" and concerned only the ARIF interface.

### *Participants*

Ten domain experts took part in the evaluation aged between twenty-eight to sixty years old. All of them were museum curators from various departments of the Victoria and Albert Museum, London, UK. No end-users were involved in the technical development of the ARCO system, so they could not be employed to assess the ARCO interface. In contrary, the museum curators were involved in the technical development from an

early stage setting user requirements and providing appropriate feedback during the early stages of implementation. They offered their knowledge about the exhibits' context and exhibitions' requirements, as well as about the visitors' needs. So, they also have been employed as expert evaluators during this phase of the evaluation and have been asked to fulfill the QUIS questionnaire.

In addition to this session, four students and two usability experts of the department of Informatics of the Aristotle University of Thessaloniki, Greece, acted as museum visitors and performed a Cognitive Walkthrough through the web-based ARIF interface, provided at `http://www.arco-web.org/vmesite/V&A/VAMGallery.html`. They were asked to assess the same aspects as the museum curators, namely the multimedia presentation in ARIF, however, they proceeded further and evaluated the overall usability under a technical point of view. These opinions are subsequently also presented.

*Instrumentation*

The QUIS (Questionnaire for User Interaction Satisfaction) questionnaire (Schneiderman and Plaisant, 2005) assessed museum curators' contentment while interacting with the ACMA and ARIF interfaces by means of a 9-scale Likert scale. This questionnaire was the main instrument to record their estimations. In contrary, the empirical evaluation used no questionnaire; however the same set of questions has been set to the usability evaluators. So, a direct comparison between the assessment of the curators (domain experts) and the "users-visitors" (usability experts) could be made.

However, the ACMA interface (ARCO Content Management Application) can by definition not be accessed by museum visitors, so it is only assessed by the domain experts.

The QUIS questionnaire consists of 7 parts. Part 1 concerns general experience with ICT (Information and Communication Technologies), is not of great importance for this study, and has not been considered. Part 2 assesses the overall user reactions as regards to the evaluated system, Part 3 concerns the windows layout of the system, Part 4 the terminology used, Part 5 the learnability of the interface (how easy it is to learn), and Part 6 the system capabilities. These first 6 parts evaluated the ACMA component, while the last Part 7 of the QUIS questionnaire concerned the multimedia presentation in ARIF, so, it could directly be combined with the evaluation of the usability experts in Greece, in order to elicit more accurate results.

*Variables and Hypotheses*

It must be explicitly stated at this point, that the statistical part of this study (the quantitative part) concerns not the assessment of the value of the interface itself. There are a number of studies evaluating the ARCO system in a holistic manner, such as (Sylaiou *et al.*, 2004) and (Sylaiou *et al.*, 2006) with concrete suggestions for the improvement of the system. This study focuses merely on the comparison of the assessments of two different groups of expert evaluators, namely, the *domain experts*, who are aware of the cultural heritage domain, yet unaware of usability aspects, and the *usability experts*, who are aware of the usability aspects, yet can act only as users-visitors in a museum context.

Under this point of view, the independent variable used in this study is of nominal type (domain or usability expert), namely the group to which the particular expert belongs, and the dependent variables are the questions in the QUIS questionnaire.

Accordingly, the hypotheses of this study are as follows:

$H_o$: *There is no difference of the evaluators' assessments due to the fact that they belong to different expert groups*.

$H_a$: *There is difference of the evaluators' assessments due to the fact that they belong to different expert groups*.

## Data and Results

### Session 1: *Curators and QUIS*

The museum curators (domain experts) evaluated by means of the QUIS a number of aspects as regards both ACMA and ARIF interfaces.

Part 2 (overall user reactions)

In this part the museum curators expressed their general opinion for the ACMA tool, in terms of *semantic bipolar differentiated* expressions, which were:

Question 1: terrible / wonderful

Question 2: frustrating / satisfying

Question 3: dull / simulating

Question 4: difficult / easy

Question 5: inadequate power / adequate power

Question 6: rigid / flexible

Table 1 summarizes the descriptive statistics for this part.

In a similar manner, all descriptive statistics for all Parts have been calculated. The pending tables are presented in the Appendix.

Part 3 (windows in the ACMA tool)

In this part the museum curators expressed their opinion on various aspects of the ACMA interface, such as windows layout, readability of displayed characters, logical

Table 1

Descriptive Statistics of Part 2 – overall impression

| | *N* | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part2-Q1 | 10 | 5.00 | 9.00 | 6.9000 | 1.10050 |
| Part2-Q2 | 10 | 2.00 | 8.00 | 5.9000 | 1.79196 |
| Part2-Q3 | 10 | 1.00 | 9.00 | 5.8000 | 2.34758 |
| Part2-Q4 | 10 | 3.00 | 9.00 | 6.1000 | 2.02485 |
| Part2-Q5 | 10 | 5.00 | 8.00 | 6.8000 | 1.03280 |
| Part2-Q6 | 10 | 6.00 | 8.00 | 7.0000 | .66667 |
| Valid *N* (listwise) | 10 | | | | |

order of displayed windows etc. However, some comments were of importance, such as "The system is configured for an experienced user. Those with less experience will quite likely need more assistance than is given" and "Sometimes, the characters are hard to read especially when rotated and floating beside a 3D model. Good when it is flat and facing the reader".

Table 5 summarizes the descriptive statistics for this part.

Part 4 (terminology)

In this part the museum curators expressed their opinion on the consistency of the terminology, the relevancy to the performed job or the appropriateness of the displayed messages. Comments, such as "There is too much jargon/technical terminology used throughout the system. It is very unintuitive" and "It seems very abstract to begin with, e.g., 'Media Object'. When you have a concept to link it to, it becomes easier" indicate a debate as regards the used terminology, discussed later.

Table 6 summarizes the descriptive statistics for this part.

Part 5 (learning the ACMA tool)

In this part the museum curators expressed their opinion on how easy or encouraging was the learning of the interface, or if exploration or remembering of important features was feasible. Representative comments: "What do you mean? It would take quite a while to absorb it all" and "The tool is quite complex. It needs more time and explanation than allowed in this testing scenario. It is also a problem that language of tests did not always conform to that used in the assessment questions. This made it difficult to know whether what you had done/experienced was what was being queried".

Table 7 summarizes the descriptive statistics for this part.

Part 6 (capabilities)

In this part the museum curators expressed their opinion on system speed, reliability, ease of operation, possibility to undo actions and correction of user mistakes.

Table 8 summarizes the descriptive statistics for this part.

Part 7 (multimedia representation in ARIF)

This part concerned the ARIF interface and the museum curators assessed the presentation of multimedia elements, such as quality of still images, sound output and colors. Comments here were from "Still images were rather fuzzy here" and "Didn't hear any sound" up to "Very impressive capabilities".

Table 9 summarizes the descriptive statistics for this part.

**Session 2:** ***A Cognitive Walkthrough in ARIF***

The next session, performed at the multimedia laboratory of the Department of Informatics, AUTh., consisted of a Cognitive Walkthrough through the ARIF interface and, because no human artifact is perfect, pinpointed also a number of usability problems. However, the usability experts had a completely different view than museum curators and made some concrete statements, such as the resolution of the screen and the level of detail of the artifacts, which were more "puristic" than it could be depicted on the QUIS.

Accordingly, the usability experts have been asked to complete the same QUIS questions concerning the ARIF interface, namely only Part 7.


Fig. 3. An artifact in the AR interface.

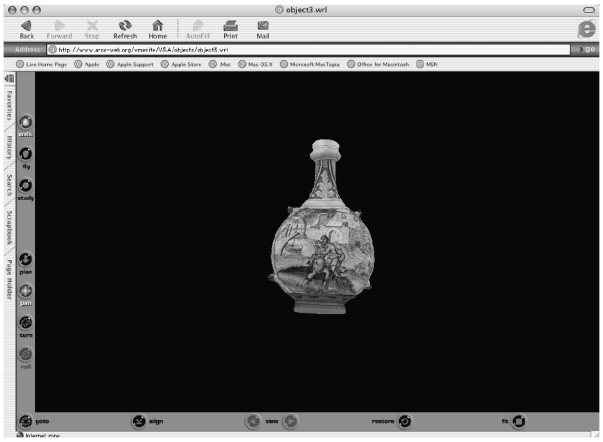

Fig. 4. Manipulation of an artifact.

## Interpretation and Discussion

As regards the first session, where only domain experts participated, the first obvious result of the above presented statistics is the low mean value of almost all questions. The highest is at 6.56, and the lowest at 5.91. Table 2 provides the mean values for all parts.

In a 9-scale Likert scale and given the relative high values people usually give in questionnaire based surveys, this is an indication of an overall “concerned acceptance” of the usability of the interface. In more detail:

The overall interface is considered to be powerful enough and flexible, although a little dull and frustrating. The handling of the various windows elements is assessed as most successful, providing the highest mean. However, the terminology used provided some scepticism. This is per se important, as museum curators are aware of the domain

Table 2
Mean values of all evaluators for all Parts of the questionnaire

| | *N* | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part 2 | 60 | 1.00 | 9.00 | 6.4167 | 1.61866 |
| Part 3 | 100 | 3.00 | 9.00 | 6.5600 | 1.20034 |
| Part 4 | 185 | 2.00 | 9.00 | 5.9135 | 1.46075 |
| Part 5 | 119 | 2.00 | 9.00 | 6.0252 | 1.60223 |
| Part 6 | 129 | 1.00 | 9.00 | 6.3876 | 1.55779 |
| Part 7 | 72 | 3.00 | 9.00 | 6.5417 | 1.57388 |
| Valid *N* (listwise) | 52 | | | | |

terminology; so this could be an indication that the terminology used in the system did not completely adhere to the domain standards. Furthermore, the system scaffolding was not adequate, as system messages, information on user progress or error messages were considered more frustrating than helpful. The learnability of the interface was also questioned, as well as the remembering of certain system commands. Finally, the multimedia presentation has been considered by the museum curators as sound.

This last point is however one of the most debatable of this study. During the second session, the usability experts in Greece, who acted as museum visitors and visited the museum through a web browser, reported a mediocre usability of the environment and a low satisfaction as regards to the cognitive aspects of the interface. They considered the environment to be unintuitive, without adequate help to scaffold novice users and with poor level of information as regards the presented artefacts. Table 3 provides their ratings depicted by means of the QUIS questions.

The provided mean values are significant lower than those of the domain experts. So, a question arose here, namely whether the answers provided by the domain experts are in accordance to those of the usability experts. In order to clarify this emerged aspect, a post-hoc elaboration procedure has been designed: An independent samples t-test as well

Table 3
Descriptive statistics for usability experts

| | *N* | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Gr Part7 | 54 | 2.00 | 8.00 | 4.5741 | 1.81817 |
| GP7Q2 | 6 | 4.00 | 7.00 | 5.0000 | 1.54919 |
| GP7Q3 | 6 | 4.00 | 8.00 | 5.6667 | 1.50555 |
| GP7Q4 | 6 | 2.00 | 6.00 | 3.6667 | 1.50555 |
| GP7Q5 | 6 | 2.00 | 7.00 | 3.8333 | 1.72240 |
| GP7Q6 | 6 | 2.00 | 6.00 | 3.6667 | 1.36626 |
| GP7Q7 | 6 | 2.00 | 6.00 | 3.8333 | 1.83485 |
| GP7Q8 | 6 | 2.00 | 7.00 | 4.5000 | 2.07364 |
| Valid *N* (listwise) | 6 | | | | |

as the non-parametric Mann–Whitney and Wilcoxon tests have been employed, in order to compare the mean values of the estimations of the two group of experts, as shown in Table 4.

In all tests, the provided statistical significance of 0.000 (SPSS cuts the rest of the decimals, indicating a very small number) in either cases (assuming or not a homogeneity of the variance of the samples, depicted by means of the Levene's test, also present it Table 4a) means a statistical significance, at a level of a p-value lower than 0.001. So, the null hypothesis must be rejected and the alternative adopted instead, so *there is difference of the evaluators' opinions due to the fact that they belong to different expert groups*.

Having statistically confirmed the disagreement of the evaluators, the next point of discussion is why there is such a great divergence between the curators' and the usability experts' opinions, as well as which is the influence of this result on the usability evaluation itself. Some studies, such as (Karoulis *et al.*, 2000), report that usability experts are usually more rigorous than users. This explanation seems however in this case not plausible for two reasons. Firstly museum curators are also experts and should also be rigorous in their estimations, and secondly, the usability experts acted in this session as real users, who were initially enthusiastic to visit the virtual museum, yet they were at the end of the session not thus enthusiastic.

Table 4

a. Independent samples t-test

| | | Levene's test for equality of variances | | t-test for equality of means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean difference | Std. error difference | 95% confidence interval of the difference | |
| | | | | | | | | | Lower | Upper |
| Part 7 | Equal variances assumed | .917 | .340 | 6.749 | 118 | .000 | 2.08333 | .30867 | 1.47208 | 2.69459 |
| | Equal variances not assumed | | | 6.589 | 92.420 | .000 | 2.08333 | .31618 | 1.45541 | 2.71126 |

b. Mann–Whitney and Wilcoxon tests[a]

| | Part 7 |
|---|---|
| Mann–Whitney U | 689.500 |
| Wilcoxon W | 1865.500 |
| Z | −5.638 |
| Asymp. Sig. (2-tailed) | .000 |

[a] Grouping variable: expert (domain/usability)

The root of this problem is probably in the nature of the evaluation. It is a fact that both groups have been asked to assess *usability features* of the interface. So, it is more plausible to believe that in this context, the usability experts are closer to the goal than the domain experts, as they know exactly *what* and *how* has to be investigated. It also seems a plausible claim the fact that museum curators have a more or less "foggy" impression of usability and its parameters, so in this context, errato humanum est. . .

## Conclusion

The first obvious conclusion is that the usability evaluation of a museum virtual interface is possible through an expert-based approach. The museum curators are aware of many aspects on this domain and they perform adequately if they are surveyed in the correct way. Their responses lead to concrete improvements of the interface and their qualitative comments, in preliminary form already presented in (Sylaiou *et al.*, 2004) are a valuable source to improve such kind of interfaces.

However, the implication of the complete opposite thesis of the usability experts who acted as users raises many questions on some biases of the questionnaire-based surveys, as already stated. A tentative claim is that an expert-based usability evaluation cannot be performed without the participation of usability experts. However, in the described context, domain experts are also inevitable. As shown in the first 6 Parts of the evaluation, the museum curators showed an overall satisfaction on the usability of the system. The debate emerged when the usability experts considered the web-interface (the ARIF) acting as users. Here the "usability expert as a user" view was very different from the "domain expert as a curator" view.

So, final conclusion of this study is that one encounters here the limits of the expert-based interface evaluation approach: *in complex interfaces, double experts (usability and domain experts) are inevitable for reliable and valid results*. Simple experts (only domain or only usability) do not seem to perform adequately. However, the fact that such double experts are extremely rare and expensive, pinpoints the aforementioned limit of the expert-based approaches. This is of course a tentative claim, as this aspect was outside of the scope of the present study; therefore, new studies must be set up in order to validate these claims.

Under this point of view, the most promising approach to evaluate interfaces of a complex kind, such as cultural heritage virtual interfaces, is the combination of expert-based approaches, which are less resource consuming, with real user-based approaches, which are more reliable, yet very resource consuming and more difficult to set up and materialize.

**Acknowledgement**

Part of this work has been funded by the Marie Curie Actions Human resources and Mobility Marie Curie training site: Virtual Reality and computer graphics, project HPMT-CT-2001-00326.

## Appendix – Descriptive Statistics Tables

Table 5
Descriptive Statistics of Part 3 – windows

| | $N$ | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part3 Q1 | 10 | 5.00 | 9.00 | 6.8000 | 1.39841 |
| Part3 Q2 | 10 | 5.00 | 8.00 | 6.9000 | .99443 |
| Part3 Q3 | 10 | 4.00 | 9.00 | 6.9000 | 1.52388 |
| Part3 Q4 | 10 | 5.00 | 8.00 | 6.5000 | .84984 |
| Part3 Q5 | 10 | 5.00 | 8.00 | 6.3000 | .94868 |
| Part3 Q6 | 10 | 5.00 | 9.00 | 6.6000 | 1.26491 |
| Part3 Q7 | 10 | 4.00 | 8.00 | 6.6000 | 1.26491 |
| Part3 Q8 | 10 | 4.00 | 8.00 | 6.6000 | 1.07497 |
| Part3 Q9 | 10 | 5.00 | 8.00 | 6.3000 | 1.15950 |
| Part3 Q10 | 10 | 3.00 | 8.00 | 6.1000 | 1.59513 |
| Valid $N$ (listwise) | 10 | | | | |

Table 6
Descriptive Statistics of Part 4 – terminology

| | $N$ | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part4 Q1 | 10 | 3.00 | 8.00 | 6.3000 | 1.49443 |
| Part4 Q2 | 10 | 5.00 | 8.00 | 6.8000 | .91894 |
| Part4 Q3 | 10 | 6.00 | 8.00 | 7.0000 | .81650 |
| Part4 Q4 | 10 | 4.00 | 8.00 | 4.9000 | 1.28668 |
| Part4 Q5 | 10 | 3.00 | 9.00 | 5.4000 | 1.64655 |
| Part4 Q6 | 10 | 3.00 | 8.00 | 5.7000 | 1.82878 |
| Part4 Q7 | 10 | 5.00 | 8.00 | 6.9000 | .99443 |
| Part4 Q8 | 10 | 6.00 | 8.00 | 7.0000 | .66667 |
| Part4 Q9 | 10 | 2.00 | 8.00 | 5.7000 | 1.70294 |
| Part4 Q10 | 10 | 3.00 | 8.00 | 5.5000 | 1.84089 |
| Part4 Q11 | 10 | 4.00 | 7.00 | 5.5000 | .97183 |
| Part4 Q12 | 9 | 3.00 | 7.00 | 5.7778 | 1.20185 |
| Part4 Q13 | 10 | 4.00 | 7.00 | 5.5000 | 1.08012 |
| Part4 Q14 | 10 | 4.00 | 8.00 | 6.3000 | 1.15950 |
| Part4 Q15 | 9 | 2.00 | 7.00 | 5.2222 | 1.39443 |
| Part4 Q16 | 10 | 4.00 | 8.00 | 6.3000 | 1.82878 |
| Part4 Q17 | 9 | 2.00 | 7.00 | 5.5556 | 1.50923 |
| Part4 Q18 | 9 | 2.00 | 6.00 | 4.8889 | 1.36423 |
| Part4 Q19 | 9 | 3.00 | 8.00 | 5.8889 | 1.53659 |
| Valid $N$ (listwise) | 8 | | | | |

Table 7
Descriptive statistics of Part 5 – learnability

| | *N* | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part5 Q1 | 10 | 3.00 | 8.00 | 5.8000 | 1.61933 |
| Part5 Q2 | 10 | 2.00 | 8.00 | 5.3000 | 2.00278 |
| Part5 Q3 | 10 | 3.00 | 8.00 | 5.7000 | 1.70294 |
| Part5 Q4 | 9 | 4.00 | 9.00 | 5.8889 | 1.69148 |
| Part5 Q5 | 10 | 3.00 | 8.00 | 6.2000 | 1.54919 |
| Part5 Q6 | 10 | 4.00 | 8.00 | 6.4000 | 1.42984 |
| Part5 Q7 | 10 | 3.00 | 8.00 | 6.0000 | 1.69967 |
| Part5 Q8 | 10 | 2.00 | 8.00 | 5.4000 | 1.95505 |
| Part5 Q9 | 10 | 2.00 | 8.00 | 5.7000 | 1.94651 |
| Part5 Q10 | 10 | 3.00 | 9.00 | 6.3000 | 1.76698 |
| Part5 Q11 | 10 | 2.00 | 8.00 | 6.3000 | 1.82878 |
| Part5 Q12 | 10 | 5.00 | 8.00 | 6.6000 | .96609 |
| Part5 Q13 | 10 | 3.00 | 8.00 | 6.0000 | 1.33333 |
| Valid *N* (listwise) | 9 | | | | |

Table 8
Descriptive statistics of Part 6 – capabilities

| | *N* | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part6 Q1 | 10 | 5.00 | 9.00 | 6.9000 | 1.28668 |
| Part6 Q2 | 10 | 4.00 | 9.00 | 6.8000 | 1.39841 |
| Part6 Q3 | 10 | 5.00 | 9.00 | 7.0000 | 1.24722 |
| Part6 Q4 | 10 | 6.00 | 8.00 | 6.9000 | .73786 |
| Part6 Q5 | 10 | 5.00 | 8.00 | 6.7000 | .94868 |
| Part6 Q6 | 10 | 5.00 | 9.00 | 7.0000 | 1.33333 |
| Part6 Q7 | 10 | 1.00 | 7.00 | 5.0000 | 1.76383 |
| Part6 Q8 | 10 | 4.00 | 8.00 | 5.8000 | 1.03280 |
| Part6 Q9 | 10 | 4.00 | 9.00 | 7.0000 | 1.56347 |
| Part6 Q10 | 10 | 3.00 | 9.00 | 5.6000 | 1.89737 |
| Part6 Q11 | 10 | 7.00 | 9.00 | 7.6000 | .69921 |
| Part6 Q12 | 9 | 1.00 | 8.00 | 5.3333 | 2.23607 |
| Part6 Q13 | 10 | 3.00 | 7.00 | 5.3000 | 1.25167 |
| Valid *N* (listwise) | 9 | | | | |

Table 9
Descriptive statistics of Part 7 – multimedia in ARIF

| | $N$ | Minimum | Maximum | Mean | Std. deviation |
|---|---|---|---|---|---|
| Part7 Q1 | 10 | 3.00 | 9.00 | 6.5000 | 1.90029 |
| Part7 Q2 | 10 | 5.00 | 9.00 | 6.8000 | 1.31656 |
| Part7 Q3 | 10 | 5.00 | 9.00 | 6.8000 | 1.39841 |
| Part7 Q4 | 8 | 4.00 | 9.00 | 6.6250 | 1.99553 |
| Part7 Q5 | 8 | 4.00 | 9.00 | 6.6250 | 2.13391 |
| Part7 Q6 | 7 | 5.00 | 9.00 | 6.7143 | 1.60357 |
| Part7 Q7 | 10 | 4.00 | 7.00 | 6.0000 | 1.15470 |
| Part7 Q8 | 9 | 4.00 | 8.00 | 6.3333 | 1.50000 |
| Valid $N$ (listwise) | 6 | | | | |

## References

ARCO, Augmented Representation of Cultural Objects (2004). `http://www.arco-web.org/` (retrived 7 Oct 2004).

Cosmas, J., T. Itegaki, D. Green (2001). 3D MURALE: a multimedia system for archaeology. In *Proceedings of the ACM-SIGGRAPH Conference on Virtual Reality, Archaeology and Cultural Heritage VAST 2001*, Athens, Greece. pp. 297–305.

Gleue, T., and P. Dähne (2001). Design and implementation of a mobile device for outdoor augmented reality in the ARCHEOGUIDE project. In *Proceedings of the Virtual Reality, Archaeology, and Cultural Heritage International Symposium* (*VAST01*), Athens, Greece.

ISO 9241 – International Standarization Organization (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDT's)*.

Jones, J., and M. Christal (2002). *The Future of Virtual Museums: On-Line, Immersive, 3D Environments*. Created Realities Group.

Karat, C., R. Campbell and T. Fiegel (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of ACM CHI '92*. Monterey, CA, May 3–7. pp. 397–404.

Lewis, C., and J. Rieman (1994). *Task-centered User Interface Design – A practical Introduction*. `ftp.cs.colorado.edu/pub/cs/distribs/HCI-Design-Book` (retrived 20 May 2001).

Liarokapis, F., N. Mourkoussis, M. White, J. Darcy, M. Sifniotis, P. Petridis, A. Basu and P.F. Lister (2004). Web3D and augmented reality to support engineering education. *World Transactions on Engineering and Technology Education*, *UICEE*, **3**(1), 11–14.

Maier, N.R.F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, **12**, 181–194.

Mourkoussis, N., M. White, M. Patel, J. Chmielewski and K. Walczak (2003). AMS – metadata for cultural exhibitions using virtual reality. In *Proc. Dublin Core Conference* (*DC2003*), Seattle, Washington.

Nielsen, J. (1993). *Usability Engineering*. Academic Press, San Diego.

Patel, M., K. Walczak, M. White and W. Cellary (2003). Digitisation to presentation-building virtual museum exhibitions. In *Proceedings of the International Conference on Vision, Video and Graphics*, Bath, UK. pp. 189–196.

Perrot, P. (1977). *The Smithsonian Experience*. Smithsonian Institute.

Preece, J., Y. Rogers, H. Sharp, D. Benyon, S. Holland, T. Carey (1994). *Human-Computer Interaction*. Addison-Wesley Publ., London.

Reeves, T.C. (1993). Evaluating technology-based learning. In G.M. Piskurich (Ed.), *The ASTD Handbook of Instructional Technology*. McGraw-Hill, New York. pp. 15.1–15.32.

Scali, G., M. Segbert, B. Morganti (2002). Multimedia applications for innovation in cultural heritage: 25 European trial projects and their accompanying measure TRIS. In *68th IFLA Council and General Conference*, August 18–24.

Schneiderman, B., and C. Plaisant (2005). *Designing the User Interface*: *Strategies for Effective Human-*

*Computer Interaction*. 4th Edition. Addison-Wesley.

Scriven, M. (1976). The methodology of evaluation. In R. Tyler (Ed.), *Perspectives of Curriculum evaluation*. Rand McNally, Chicago.

Shaw, N., M. Spearman, J. Hemsley and M. Kaayuk (2004). *Report on Current Practices and Needs*. ORION (Object Rich Information Network), Deliverable 5.
`http://www.orion-net.org/orion_library.asp` (retrived 7 Oct 2004).

Sylaiou S., A. Almosawi, K. Mania, M. White (2004). Preliminary evaluation of the augmented representation of cultural objects system. In *Proceedings of the 10th International Conference on Virtual Systems and Multimedia, Hybrid Realities-Digital Partners, Explorations in Art, Heritage, Science and the Human Factor*, 17–19 November, Softopia Japan, Ogaki City, Japan. pp. 426–431.

Sylaiou, S., A. Karoulis and M. White (2006). Evaluation of the augmented representation of a cultural objects system (submitted).

Tsapatori, M. (2003). *ORION Research Roadmap, Evaluation and Assessment*. Object Rich Information Network (ORION), Deliverable 8.
`http://www.orion-net.org/Admin/LibraryLoc/file28.pdf` (retrieved 20 April 2005).

Wojciechowski, R., K. Walczak, M. White and W. Cellary (2004). Building virtual and augmented reality museum exhibitions. In *Proceedings of the Web3D 2004 Symposium – the 9th International Conference on 3D Web Technology*, *ACM SIGGRAPH*, Monterey, California (USA). pp. 135–144.

**A. Karoulis**

**S. Sylaiou**

**M. White**